Table 1: Qwen-2.5-32B-Instruct pruned to 2:4 sparsity with Wanda.

Cal. Set	Block size	PPL	Acc.	HumanEval	GSM8K
C4	Per matrix	14.42	66.84%	21.51%	65.19%
C4	1/2	6.39	69.04%	26.22%	72.56%
C4	1	6.40	69.10%	25.87%	73.31%
C4	8	6.37	69.08%	26.14%	73.26%
OpenCode	Per matrix	15.26	61.85%	53.02%	71.13%
OpenCode	1/2	8.11	65.80%	54.24%	73.96%
OpenCode	1	8.08	65.88%	54.87%	74.14%
OpenCode	8	8.02	65.77%	54.52%	74.09%
GSM8K	Per matrix	14.85	61.62%	44.01%	75.52%
GSM8K	1/2	8.72	64.18%	47.76%	79.77%
GSM8K	1	8.68	64.24%	48.17%	79.98%
GSM8K	8	8.71	64.25%	48.12%	79.93%

Table 2: Qwen-2.5-32B-Instruct pruned to 2:4 sparsity with magnitude pruning.

Cal. Set	Block size	PPL	Acc.	HumanEval	GSM8K
C4	Per matrix	23.42	64.61%	15.10%	52.92%
C4	1/2	6.42	69.11%	26.07%	72.44%
C4	1	6.41	68.93%	26.03%	73.04%
C4	8	6.35	69.10%	25.96%	73.01%
OpenCode	Per matrix	26.91	63.27%	33.02%	54.09%
OpenCode	1/2	8.11	65.85%	54.21%	74.11%
OpenCode	1	8.08	65.73%	54.43%	74.09%
OpenCode	8	8.02	65.89%	54.22%	74.12%
GSM8K	Per matrix	25.14	59.63%	27.14%	57.94%
GSM8K	1/2	8.84	63.96%	47.82%	79.81%
GSM8K	1	8.81	64.12%	48.03%	79.94%
GSM8K	8	8.84	64.08%	47.96%	79.87%

Table 3: LLaMA-3-8B pruned to 2:4 sparsity with Wanda.

Cal. Set	Block size	PPL	Acc.	HumanEval	GSM8K
C4	Per matrix	18.81	48.26%	3.27%	5.22%
C4	1/2	10.02	55.52%	5.94%	10.38%
C4	1	10.19	55.69%	6.09%	10.06%
C4	8	10.85	54.40%	5.92%	9.89%
OpenCode	Per matrix	19.93	46.71%	9.04%	8.15%
OpenCode	1/2	14.08	51.88%	20.82%	15.21%
OpenCode	1	13.89	52.71%	21.31%	15.38%
OpenCode	8	14.83	51.15%	20.07%	14.46%
GSM8K	Per matrix	19.39	45.04%	3.82%	12.26%
GSM8K	1/2	15.89	48.47%	7.09%	21.85%
GSM8K	1	15.56	50.61%	6.48%	23.49%
GSM8K	8	15.28	49.27%	7.31%	22.44%

Table 4: LLaMA-3-8B pruned to 2:4 sparsity with magnitude pruning.

Cal. Set	Block size	PPL	Acc.	HumanEval	GSM8K
C4	Per matrix	45.81	44.60%	2.99%	3.14%
C4	1/2	12.73	53.39%	4.16%	5.90%
C4	1	12.97	53.76%	4.04%	5.77%
C4	8	13.14	52.43%	4.26%	5.91%
OpenCode	Per matrix	68.55	39.14%	7.31%	7.08%
OpenCode	1/2	26.29	46.09%	12.17%	10.12%
OpenCode	1	24.75	45.04%	12.29%	10.67%
OpenCode	8	25.38	44.71%	11.85%	10.26%
GSM8K	Per matrix	56.31	41.12%	2.51%	8.25%
GSM8K	1/2	18.67	48.36%	4.49%	18.43%
GSM8K	1	20.17	48.06%	4.52%	18.14%
GSM8K	8	19.31	46.92%	4.34%	17.29%

Table 5: LLaMA-3-8B pruned to 2:4 sparsity with Wanda and reconstructed with different block sizes and calibration set sizes.

Cal. set size / Block size	16	32	64	128	256	512	1024	2048	4096	8192
Per matrix	20.85	19.99	19.37	18.86	18.81	18.77	18.80	18.79	18.80	18.76
1/2	15.43	14.07	13.69	13.07	12.16	11.51	11.24	11.18	10.98	10.78
1	14.78	13.70	13.33	12.77	12.06	11.49	11.29	11.25	10.95	10.82
8	15.11	13.84	13.36	12.81	12.09	11.54	11.26	11.20	10.92	10.84

Table 6: Reconstruction error after transformer blocks of LLaMA-3-8B pruned to 2:4 sparsity with Wanda and reconstructed at different granularities.

Transformer block / Block size	1	4	8	12	16	20	24	28	32
Per matrix	0.14	5.20	9.77	14.86	20.21	36.65	68.05	123.46	166.82
1/2	0.05	3.46	7.16	10.43	15.84	28.84	52.99	92.84	118.02
1	0.05	3.89	7.59	10.99	16.56	29.62	53.80	94.32	120.45
2	0.07	4.01	7.71	11.34	16.52	30.32	53.80	92.70	123.21
8	0.09	3.92	7.62	11.43	16.50	29.94	54.06	93.80	120.79

Table 7: LLaMA-3-8B pruned to different sparsities with Wanda and reconstructed at different block sizes.

Sparsity / Block size	0.4	0.5	0.6	0.7	0.8	0.9
No rec.	7.37	9.12	18.11	87.93	691.64	18693.60
Per matrix	7.28	8.78	15.32	21.91	66.22	3501.47
1/2	7.13	8.08	10.18	16.85	60.16	2490.68
1	7.19	8.14	11.14	16.65	59.70	1469.93
2	7.15	8.18	10.16	17.90	59.68	2866.76
8	7.20	8.26	10.86	19.20	60.35	1272.12

Table 8: LLaMA-3-8B pruned to different sparsities with magnitude pruning and reconstructed at different block sizes.

Sparsity / Block size	0.4	0.5	0.6	0.7	0.8	0.9
No rec.	32.25	96.12	352.48	847.28	2.64×10^5	5.82×10^6
Per matrix	12.18	19.64	31.33	54.52	76623.15	3.57×10^5
1/2	7.39	8.71	11.97	22.98	46267.20	37468.30
1	8.62	9.81	12.57	21.34	21348.80	13152.60
2	7.53	8.81	11.95	21.21	44843.70	2.34×10^5
8	7.65	9.41	12.59	24.56	123.04	21685.00

Table 9: LLaMA-3-8B pruned to 20% and 50% structured sparsity and reconstructed at different granularities.

Block size	Method	Sparsity	PPL no rec.	PPL rec.	Acc. no rec.	Acc rec.
1	FLAP	20%	7.81	7.03	53.42%	57.68%
1	Olica	20%	7.25	6.98	54.11%	57.92%
1	FLAP	50%	46.25	17.51	42.87%	48.76%
1	Olica	50%	32.27	16.84	44.27%	49.14%
8	FLAP	20%	7.81	7.09	53.42%	57.52%
8	Olica	20%	7.25	7.04	54.11%	57.86%
8	FLAP	50%	46.25	17.18	42.87%	48.83%
8	Olica	50%	32.27	16.91	44.27%	49.08%

Table 10: LLaMA-3-8B pruned to 2:4 sparsity with Wanda and only partially reconstructed. For block size 8, we do not update the parameters of skipped blocks while reconstructing the eight block submodel submodel that contains them.

Blocks skipped/ Block size	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	No rec.
Per matrix	18.81	18.94	19.02	19.08	19.13	19.48	19.75	20.10	20.63	21.27	21.58
1/2	10.02	10.31	10.51	10.58	10.75	11.90	13.27	15.45	19.22	20.28	21.58
1	10.19	10.46	10.50	10.60	10.77	11.87	13.30	15.49	19.20	20.32	21.58
8	10.85	10.98	11.04	11.12	11.19	11.92	13.63	16.08	19.84	20.79	21.58